LLNL-JRNL-407403

An Extended Study of the Discriminant Random Forest

T. D. Lemmond, B. Y. Chen, A. O. Hatch, W. G. Hanley

October 1, 2008

Annals of Information Systems

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

An Extended Study of the Discriminant Random Forest

Tracy D. Lemmond, Barry Y. Chen, Andrew O. Hatch and William G. Hanley

Systems and Decision Sciences, Lawrence Livermore National Laboratory

Abstract Classification technologies have become increasingly vital to information analysis systems that rely upon collected data to make predictions or informed decisions. Many approaches have been developed, but one of the most successful in recent times is the Random Forest. The Discriminant Random Forest is a novel extension of the Random Forest classification methodology that leverages Linear Discriminant Analysis to perform multivariate node splitting during tree construction. An extended study of the Discriminant Random Forest is presented which shows that its individual classifiers are stronger and more diverse than their Random Forest counterparts, yielding statistically significant reductions in classification error of up to 79.5%. Moreover, empirical tests suggest that this approach is computationally less costly with respect to both memory and efficiency. Further enhancements of the methodology are investigated that exhibit significant performance improvements and greater stability at low false alarm rates.

1 Introduction

One of the greatest emerging assets of the modern technological community is *information*, as the computer age has enhanced our ability to collect, organize, and analyze large quantities of data. Many practical applications rely upon systems that are designed to assimilate this information, enabling complex analysis and inference. In particular, classification technologies have become increasingly vital to systems that learn patterns of behavior from collected data to support prediction and informed decision-making. Applications that benefit greatly from these methodologies span a broad range of fields, including medical diagnostics, network analysis (e.g., social, communication, transportation, and computer networks), image analysis, natural language processing (e.g., document classification), speech recognition, and numerous others.

Many effective approaches to classification have been developed, but one of the most successful in recent times is the Random Forest. The Random Forest (RF) is a nonparametric ensemble classification methodology whose class predictions are based upon the aggregation of multiple decision tree classifiers. In this paper, we present an in-depth study of the Discriminant Random Forest (DRF), a novel classifier that extends the conventional RF via a multivariate node splitting technique based upon a linear discriminant function.

Application of the DRF to various two-class signal detection tasks has demonstrated that this approach achieves reductions in classification error of up to 79.5% relative to the RF. Empirical tests suggest that this performance improvement can be largely attributed to the enhanced strength and diversity of its base tree classifiers, which, as demonstrated in [3], lead to lower bounds on the generalization error of ensemble classifiers. Moreover, experiments suggest that the DRF is computationally less costly with respect to both memory and efficiency.

This paper is organized as follows: Section 2 summarizes the motivation and theory behind the Random Forest methodology. We present the Discriminant Random Forest approach in detail in Section 3, and contrast the performance of the RF and DRF methods for two signal detection applications in Section 4. This study incorporates an assessment of statistical significance of the observed differences in algorithm performance. Finally, our conclusions are summarized in Section 5.

2 Random Forests

The random decision forest concept was first proposed by Tin Kam Ho of Bell Labs in 1995 [12, 13]. This method was later extended and formalized by Leo Breiman, who coined the more general term *Random Forest* to describe the approach [3].

In [3], Breiman demonstrated that RFs are not only highly effective classifiers, but they readily address numerous issues that frequently complicate and impact the effectiveness of other classification methodologies leveraged across diverse application domains. In particular, the RF requires no simplifying assumptions regarding distributional models of the data and error processes. Moreover, it easily accommodates different types of data and is highly robust to overtraining with respect to forest size. As the number of trees in the RF increases, the generalization error, PE^*, has been shown in [3] to converge and is bounded as follows,

$$PE^* \leq \frac{\overline{\rho}(1-s^2)}{s^2} \tag{1}$$

$$s = 1 - 2 \cdot PE^*_{tree} \tag{2}$$

where $\overline{\rho}$ denotes the mean correlation of tree predictions, s represents the strength of the trees, and PE^*_{tree} is the expected generalization error for an individual tree classifier. From (Eq. 1), it is immediately apparent that the bound on generalization error decreases as the trees become stronger and less correlated. To reduce the mean correlation, $\overline{\rho}$, among trees, Breiman proposed a bagging approach [1, 2], in which each tree is trained on a bootstrapped sample of the original training data, typically referred to as its bagged training set [5, 9]. Though each bagged training set contains the same number of samples as the original training data, its samples are randomly selected with replacement and are representative of approximately $\frac{2}{3}$ of the original data. The remaining samples are generally referred to as the *out-of-bag* (OOB) data and are frequently used to evaluate classification performance.

At each node in a classification tree, *m* features are randomly selected from the available feature set, and the single feature producing the “best” split (according to some predetermined criterion, e.g., Gini impurity) is used to partition the training data. As claimed in [3], small values of *m*, relative to the total number of features, are often sufficient for the forest to approach its optimal performance. In fact, large values of *m*, though they may increase the strength of the individual trees, induce higher correlation among them, potentially reducing the overall effectiveness of the forest. The quantity *m* is generally referred to as the *split dimension*.

Each tree is grown without pruning until the data at its leaf nodes are homogeneous, or until some other predefined stopping criterion is satisfied. Class predictions are then performed by propagating a test sample through each tree and assigning a class label, or *vote*, based upon the leaf node that receives the sample. Typically, the sample is assigned to the class receiving the majority vote. Note, however, that the resulting votes can be viewed as approximately i.i.d. random variables, and thus, the Laws of Large Numbers imply that their empirical frequency will approach their true frequency as the number of trees increases. Moreover, the empirical distribution function from which they are drawn will converge to the true underlying distribution function [14]. Ultimately, we can treat the resulting vote frequencies as class-specific probabilities and threshold upon this distribution to make a classification decision.

3 Discriminant Random Forests

The following section describes the Discriminant Random Forest methodology in greater detail. As an ensemble classification method that utilizes bagging and randomized feature selection, the DRF shares the same theoretical foundation that affords the RF its remarkable effectiveness. However, the key distinction between these techniques lies in the prescribed method for splitting tree nodes. The DRF leverages the parametric multivariate discrimination technique called Linear Discriminant Analysis.

3.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), pioneered by R.A. Fisher in 1936, is a discrimination technique that utilizes dimensionality reduction to classify items into distinct groups [8, 11, 16]. The LDA is an intuitively appealing methodology that makes class assignments by determining the linear transformation of the data in feature space that maximizes the ratio of their between-class variance to their within-class variance, achieving the greatest class separation, as illustrated in Fig. 1. The result is a linear decision boundary, identical to that determined by maximum likelihood discrimination, which is optimal (in a Bayesian sense) when the underlying assumptions of multivariate normality and equal covariance matrices are satisfied [15]. It can be shown that, in the two-class case, the maximum class separation occurs when the vector of coefficients, **w**, and intercept, *b*, used to define the linear transformation are as follows

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_0) \,, \tag{3}$$

$$b = -0.5 * (\mu_1 + \mu_0)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_0) + \log(\frac{\pi_1}{\pi_0}) \tag{4}$$

where Σ is the common covariance matrix, μ_k is the mean vector for class *k* and π_k is the prior probability of the *k*th class. Typically, when data are limited, we estimate Σ with the pooled covariance estimate, $\mathbf{S}_W$, given by

$$\mathbf{S}_W = \sum_{k=1}^{N} \mathbf{S}_k \,, \tag{5}$$

$$\mathbf{S}_k = \sum_{i=1}^{N_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)^T \,. \tag{6}$$

In the above equations, $\mathbf{x}_{ki}$ and $\bar{\mathbf{x}}_k$ denote the *i*th training sample of class *k* and the corresponding class sample mean, respectively.

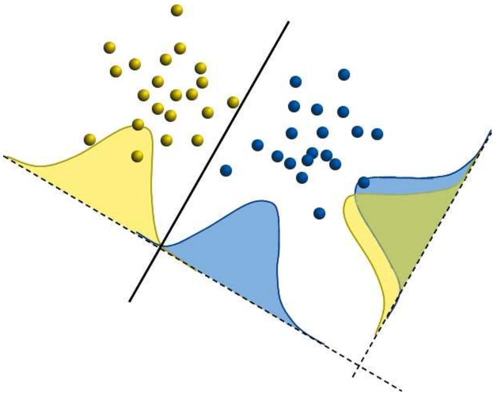

Fig. 1. LDA transformation and the optimal linear decision boundary.

3.2 The Discriminant Random Forest Methodology

Numerous variations of the Random Forest methodology have been proposed and documented in the literature, most of which address node-splitting techniques [4, 6]. Many of these are based upon an assessment of *node impurity* (i.e., heterogeneity) and include entropy-based methods, minimization of the Gini impurity index, or minimization of misclassification errors. Additional forest-based methods that focus upon alternative aspects of the algorithm include supplementing small feature spaces with linear combinations of available features [3], variations on early stopping criteria, selecting the split at random from the *n* best splits [6], and PCA transformation of random feature subsets [17].

Our Discriminant Random Forest is a novel approach to the construction of a classification tree ensemble in which LDA is employed to split the feature data. Bagging and random feature selection are preserved in this approach, but unlike other forest algorithms, we apply LDA to the data at each node to determine an "optimal" linear decision boundary. By doing so, we allow decision hyperplanes of any orientation in multidimensional feature space to separate the data, in contrast to the conventional random forest algorithm, whose boundaries are limited to hyperplanes orthogonal to the axis corresponding to the feature yielding the best split. We have illustrated this effect in Fig. 2, which depicts decision lines in two-dimensional space for the RF (left) and the DRF (right).

These two approaches to node splitting give rise to highly distinctive decision regions. Fig. 3 shows an example of a two-dimensional decision region created by each forest, in which bright blue and bright gold areas represent regions of high posterior probability for the positive and negative classes, respectively. Darker areas indicate regions of greater uncertainty. The decision region produced by the DRF is notably more complex, and its boundaries are fluid and highly intricate, fitting more closely to the training data.

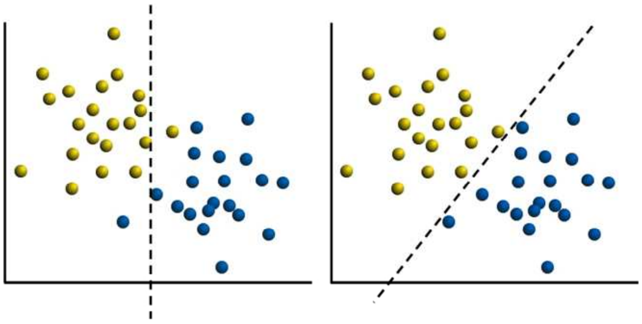

Fig. 2. Single feature splitting (left); LDA splitting (right)

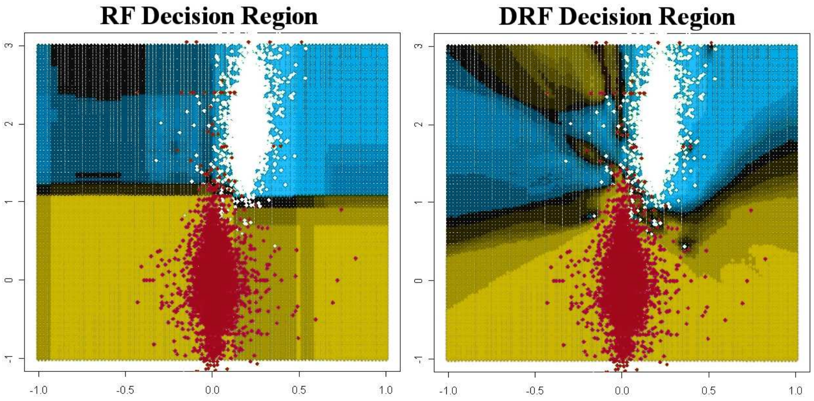

Fig. 3. Posterior probability (blue/gold represent the positive/negative classes) for the RF and DRF; the training set is overlaid, where white/maroon represent the positive/negative samples.

```
Train_DRF (Data, $m$, $NumTrees$):
  for (i = 0; i < $NumTrees$; i++)
      $D_i$ = bootstrap sample of size $N$ from Data
      Train_DRF_Tree ($D_i$,$m$)
  end for
end Train_DRF
```

```
Train_DRF_Tree($D$,$m$):
  level = 0
  create root node at level with data $D$
  while not (all nodes at level are terminal)
    for (non-terminal node $j$ at level)
      $F_j$ = sample $m$ features w/o replacement
      $D_j'$ = project $D_j$ onto $F_j$
      compute $w_j$ and $b_j$ using $D_j'$; store in $j$
      $D_L, D_R$ = split $D_j$ such that
          $D_L = x_j$ if $w_j^T x_j' + b_j > 0, \forall x_j \in D_j, x_j' \in D_j'$
          $D_R = x_j$ otherwise
      create left_child at level+1 with $D_L$
      if ($D_L$ is homogeneous)
         assign class($D_L$) to left_child
      end if
      create right_child at level+1 with $D_R$
      if ($D_R$ is homogeneous)
         assign class($D_R$) to right_child
      end if
    end for
  end while
  increment level
end Train_DRF_Tree
```

Fig. 4. Pseudocode for the Discriminant Random Forest algorithm.

Pseudocode for training a DRF on a data set of size N is provided in Fig. 4. As previously discussed, the growth of the forest proceeds in a manner similar to the conventional Random Forest, with the notable exception of the decision boundary computation, which proceeds as described in Section 3.1. Note that the termination criterion for the leaf nodes in the given algorithm relies upon homogeneity of the node data. An alternative to this approach will be presented and discussed in Section 4.4.

4 DRF and RF: An Empirical Study

In the following suite of experiments, we compare the classification performance of the RF (utilizing the misclassification minimization node-splitting criterion) and DRF for two signal detection applications: (1) detecting hidden signals of varying strength in the presence of background noise, and (2) detecting sources of radiation. Both tasks represent two-class problems, and like many other real-world applications, the costs for distinct types of error are inherently unequal. Thus, we evaluate the performance of the RF and DRF methodologies in terms of *false positives*, also known as false alarms or type I errors, and *false negatives*, also known as misses or type II errors. The false alarm rate (*FAR*), false negative rate (*FNR*), and true positive rate (*TPR* or detection rate) are defined as follows:

$$
\begin{aligned}
FAR &= \frac{\#\, negative\ samples\ misclassified}{\#\, negative\ samples} \\
FNR &= \frac{\#\, positive\ samples\ misclassified}{\#\, positive\ samples} \qquad (7) \\
TPR &= 1 - FNR
\end{aligned}
$$

4.1 Hidden Signal Detection

The goal in the *Hidden Signal Detection* application is to detect the presence of an embedded signal. In this application, it is assumed that each detection event requires a considerable amount of costly analysis, making false alarms highly undesirable. Hence, we have computed the *Area Under the Receiver Operating Characteristic (ROC) Curve*, or *AUC*, integrated over the *FAR* interval [0, 0.001] and scaled so that a value of 100% represents a perfect detection rate over this low *FAR* interval. The resulting quantity provides us with a single value that can be used to compare the prediction performance of the classifiers.

The data for these experiments are composed of two separate sets. The training data set, *T1*, consists of 7931 negative class samples (i.e., no embedded signal) along with two sets of positive class samples having 40% and 100% embedded signal strength (7598 and 7869 samples, respectively). The *J2* data set contains 9978 negative class samples and five positive classes having signal strengths of 20%, 40%, 60%, 80% and 100% (7760, 9143, 9327, 9387 and 9425 samples, respectively). The training and testing data sets for each of the following experiments consist of the negative class combined with one of the available positive classes, as indicated in each case. All data samples consist of eight features useful for detecting the presence of embedded signals. We have applied both the RF and DRF forest methodologies at each split dimension $(m \in \{1,2,\ldots,8\})$ in an effort to assess the impact of this parameter on their performance.

4.1.1 Training on *T1*, Testing on *J2*

Fig. 5 shows the plots of the *AUC* generated by training the forests on *T1* and testing on *J2* at signal strengths of 40% and 100%. Each RF or DRF was composed of 500 trees, a sufficient forest size to ensure convergence in *AUC*. In 14 of the 16 possible combinations of signal strength and split dimension, the DRF performance clearly exceeded that of the RF over the *FAR* region of interest. In the remaining two cases, the difference in the detection rate was negligible. Moreover, these results suggest that the DRF is more successful than the RF algorithm in detecting weaker signals and better utilizes more input features. As the split dimension increases, we would expect the trees to become more correlated for both methodologies, resulting in poorer prediction performance. Fig. 5 suggests this trend, but the effect appears to be noticeably less severe for the DRF. The tradeoff between tree strength and correlation with respect to *m* is discussed in greater detail in Section 4.2. ROC curves for both RF and DRF for the Hidden Signal Detection application are plotted in Fig. 6, again indicating that the DRF exhibits superior performance across the low *FAR* region of interest.

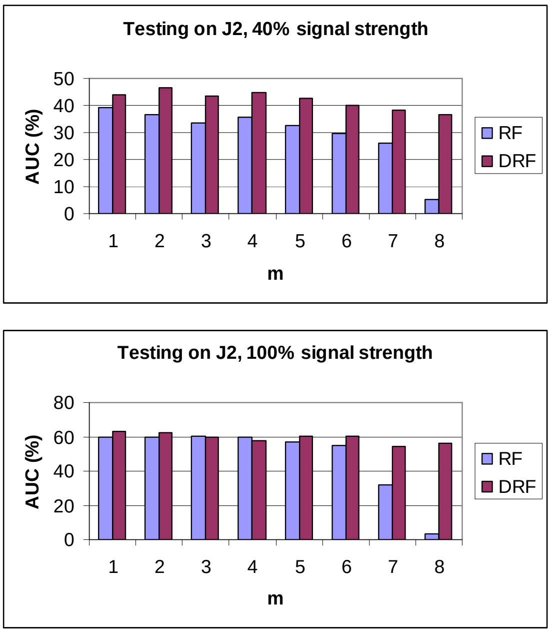

Fig. 5. AUC for RF and DRF: (top) trained on *T1*, tested on *J2* with 40% signal strength; (bottom) trained on *T1*, tested on *J2* with 100% signal strength. Each classifier is composed of 500 trees.

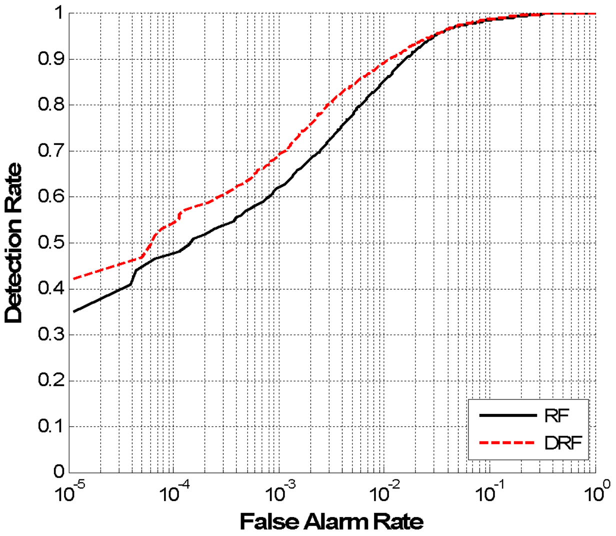

Fig. 6. ROC curves for RF *m*=1 and DRF *m*=2 trained on *T1*, tested on *J2* with 100% signal strength and 170K additional negative samples.

4.1.2 Prediction Performance for *J2* with Cross Validation

To more thoroughly explore the impact of signal strength on prediction performance, we trained and tested the RF and DRF on all signal strengths of the *J2* data set. We used 5-fold cross-validation (CV) to evaluate the performance of both classifiers. In *k*-fold cross-validation, the data set is randomly partitioned into *k* equal-sized and equally-proportioned subsets. For each run *i*, we set aside data subset *i* for testing, and we train the classifier on the remaining *k*-1 subsets. We use the average of these *k* estimates to compute our performance estimate. Fig. 7 shows the percentage increase in the *AUC* achieved by the DRF for split dimensionalities $m \in \{1,2\}$ as compared to the best-performing RF (i.e., *m*=1). As we observed when training on *T1*, the DRF yields substantially better detection performance on weaker signals.

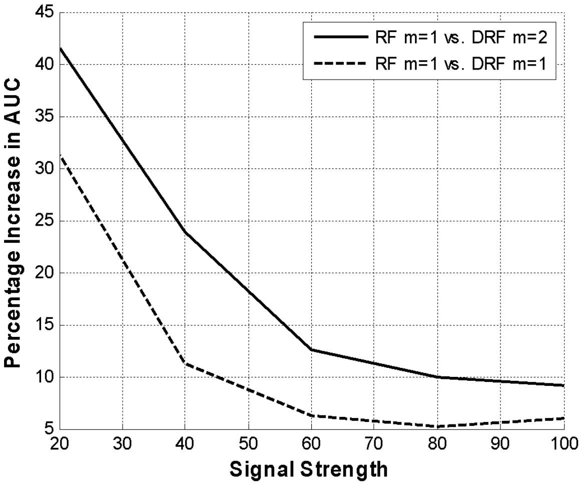

Fig. 7. Percentage increase in AUC on J2 for DRF *m* = 1, 2 over RF *m* = 1 at each signal strength. Each forest is composed of 500 trees.

4.2 Radiation Detection

The objective of the Radiation Detection effort is to detect the presence of a radiation source in vehicles traveling through a radiation portal monitoring system that measures the gamma ray spectrum of each vehicle quantized into 128 energy bins. The 128-dimensional normalized gamma ray spectra serve as the input features for the RF and DRF classifiers. The negative class is composed of radiation measurements from real vehicles containing no radiation source.

The data for the positive class were created by injecting a separate set of negative samples with spectra derived from two isotopic compositions of both Uranium and Plutonium in IAEA Category 1 quantities [18]. These sets of positive and negative samples were then partitioned into non-overlapping, equally-proportioned training and testing sets containing 17,000 and 75,000 samples, respectively.

The ROC curves in Fig. 8 show the *TPR* (i.e., detection rate) versus the *FAR* for RF and DRF. Over most of the low *FAR* range, the DRF maintains higher detection rates than the RF.

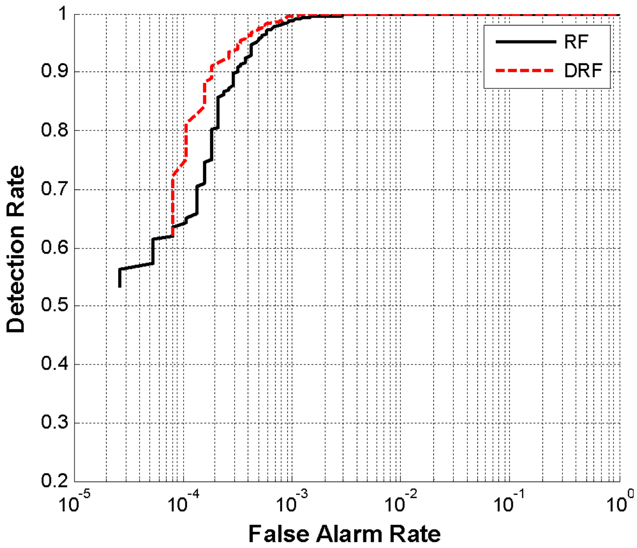

Fig. 8. ROC curves in the Radiation Detection task for RF and DRF.

The large number of features available for this application presents an ideal opportunity to thoroughly explore the behavior of the RF and DRF methodologies for high split dimensions. In particular, we wish to investigate their relative computational efficiency, memory considerations, bounds on their generalization error (Eq. 1), the interplay between tree strength and correlation, and the impact of each of these characteristics on overall algorithm performance. We have utilized the OOB data to compute these characteristics (see [3] for further details) at the minimal forest size required for each algorithm to achieve its peak performance. This can be readily observed in Fig. 9, which shows a plot of the minimum classification error (*MCE*) achieved by both classifiers with respect to $m \in \{2^n \mid n = 0,1,\ldots,7\}$. The *MCE* is plotted as a function of the forest size and indicates that the peak DRF performance was achieved by a forest consisting of approximately 50 trees, far fewer than the 250 trees required by the RF.

In Table 1, performance statistics have been provided for the RF yielding the best *MCE* and for a DRF whose performance exceeded that of the RF with respect to error, computational requirements and efficiency[1]. Both were trained on a dual-core Intel 6600 2.4 GHz processor with 4GB of RAM. To compute memory usage, we assumed that each RF node must store an integer feature ID and its corresponding floating-point threshold. Each DRF node must store *m* integers for its selected feature IDs along with *m*+1 floating point values for its weight vector, **w**.

TABLE 1. PERFORMANCE SUMMARY FOR RF / DRF.

	RF	DRF	Rel. Diff.
Dimension	4	8	------
Forest Size	250	50	80%
Avg.Nodes/Tree	1757.69	718.16	59.1%
Classification Error	4.37e-3	2.05e-3	53.0%
Training Time (s)	315	48	84.8%
Memory Usage (b)	439423	323172	26.5%

[1] With respect to *MCE* alone, the best DRF achieved a 79.5% reduction relative to the RF.

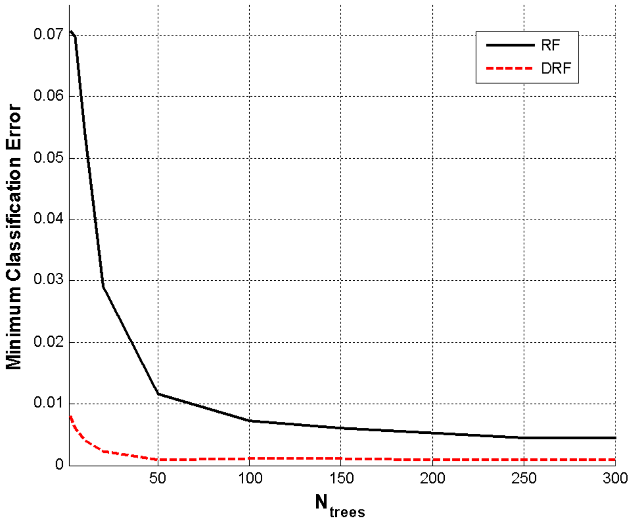

Fig. 9. Minimum classification error (MCE) for the RF and DRF as a function of forest size on the Radiation Detection dataset.

Table 1 indicates that the DRF was able to achieve a lower classification error rate than the RF while simultaneously reducing training time and memory usage by 84.8% and 26.5%, respectively. The smaller DRF trees clearly contribute to this improvement in efficiency, but their reduced size also suggests a dramatic increase in tree strength.

From an empirical standpoint, a node splitting strategy that effects a better separation of the data, such as the multivariate LDA technique, naturally generates smaller classification trees as the split dimensionality increases, as shown in Fig. 10. Though such trees might exhibit superior prediction capabilities (i.e., greater strength), we would generally expect the variation among them to decrease (i.e., increased correlation), potentially leading to a reduction in overall performance.

The key to informative analysis of these two classification methodologies, as introduced in Section 2, lies in our ability to successfully characterize this interplay between the strength and correlation of individual trees. To provide further insight into these behaviors, Fig. 11 compares the OOB estimates of tree strength and correlation for the RF and DRF, along with their classification error and respective generalization error bounds plotted as a function of the split dimensionality, *m*. As expected, the strength of an individual DRF tree is, in general, significantly greater than that of its RF counterpart. Far more remarkable is the reduced correlation among the DRF trees for split dimensions up to $m \approx 90$. The relationship between strength and correlation is typically regarded as a tradeoff [3], in which one is improved at the expense of the other, and the smaller DRF trees might naively be expected to exhibit greater correlation. However, [3] suggests that each base classifier is primarily influenced by the parameter vector representing the series of random feature selections at each node. In the multivariate setting, the number of potential feature subsets at each node increases combinatorially, dramatically enhancing the variability in the parameter vector that characterizes the classifier, which may explain the immediate drop in correlation as *m* increases. As *m* approaches the cardinality of the feature set, however, we observe a sudden and severe rise in correlation, behavior that is consistent with the reduced variation in the nodal features used for splitting.

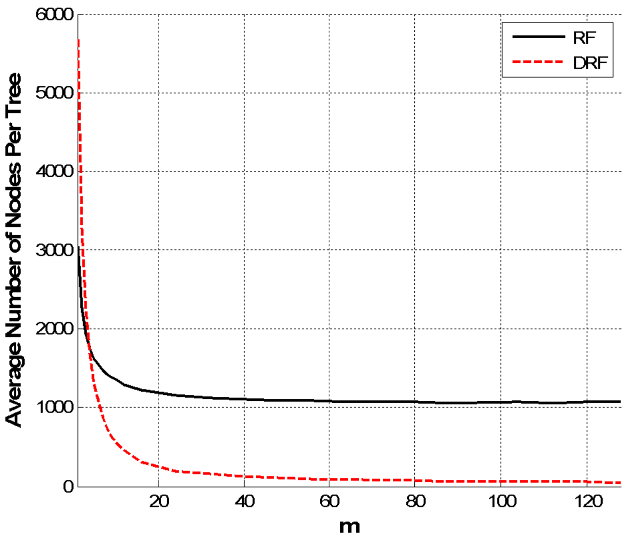

Fig. 10. Average decision tree size for the RF and DRF as a function of the dimensionality, *m*.

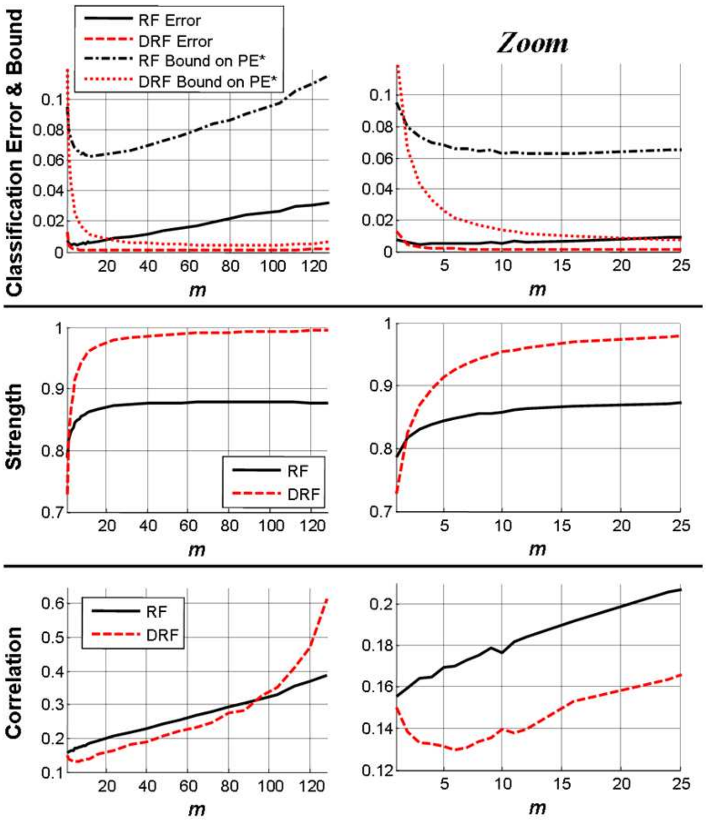

Fig. 11. OOB statistics as a function of the dimensionality, m, on the Radiation Detection data set.

Consistent with the generalization error bound (Eq. 1), Fig. 11 shows that strength has a greater impact on the classifier performance than the correlation. Even at the highest split dimensionality, the DRF classification error and error bound exhibit only minimal degradation. In contrast, the RF error steadily increases with the split dimensionality. Moreover, the DRF bound is far tighter than that of the RF, even surpassing the RF classification error at $m \approx 25$.

Interestingly, though the strength and correlation of both methodologies exhibit similar trends, their classification errors exhibit opposing behavior, suggesting that the relationship between strength and correlation is more complex than can be fully explained by our initial experiments.

4.3 Significance of Empirical Results

For the two signal detection applications discussed above, the performance of the DRF and RF methodologies was compared and contrasted via a series of experiments. The empirical evidence presented indicated that the DRF outperformed the RF with respect to each of the identified measures of performance. However, a natural question arises: *Are these observed differences statistically significant or simply an artifact of random fluctuations originating from the stochastic nature of the algorithms (e.g., bootstrap sampling of data and features)?* To more thoroughly investigate this issue, we revisited the Hidden Signal Detection application introduced in Section 4.1. Specifically, using the original *T1* training data set and a new testing set called *J1*, we wish to build statistical confidence regions surrounding "average" DRF and RF ROC curves. The resultant confidence regions could then be used to determine whether the observed differences in performance are significant.

The *J1* data set is statistically equivalent to the original data set, *J2*. It contains 179,527 negative class samples and 9,426 positive class samples with 100% embedded signal strength. As in the prior Hidden Signal Detection experiments, all data samples consist of eight features useful for detecting the presence of embedded signals.

The *J1* data set, though equivalent to *J2*, was not utilized in any fashion during the development of the DRF algorithm; consequently, it is an ideal testing data set for independently assessing the performance of the methodology. This is common practice in the speech recognition field and the broader machine learning community and is employed to prevent the subtle tuning of a methodology to a particular set of testing data.

For this study, we applied both the RF and DRF methodologies at all split dimensions $m = \{1,2,3,...,8\}$ and observed that the optimal RF occurred at $m = 2$, while the performance of the DRF peaked for $m = 1$. We will focus on these cases for the remainder of this discussion.

For each algorithm, 101 classifiers were trained and tested using variable random seeds. Based upon the resulting ROC curves, a “median” ROC and corresponding upper and lower confidence limits were computed for each methodology using a variant of the vertical averaging approach described by Fawcett [10]. Specifically, for each FAR value $\alpha \in [0.0,1.0]$, the 101 corresponding detection rates were ranked, and their median detection rate $MDR(\alpha)$ was computed along with their 97.5 and 2.5 percentiles. Using this data, the median ROC, consisting of the collection of points $\{(\alpha, MDR(\alpha) : \alpha \in [0.0,1.0]\}$, and the 97.5 and 2.5 percentile bands were computed and are shown in Fig. 12.

It is immediately apparent that the median ROC and associated percentile bands for the DRF and RF do not overlap for FAR values less than 10^{-3}, providing considerable evidence that the observed performance improvement exhibited by the DRF over this region is statistically significant.

We can also examine these results from the perspective of the AUC. Specifically, we computed the AUC values for the 101 classification results of each methodology using FAR cutoff levels of 10^{-3}, 10^{-4} and 10^{-5}. Box plots of these results, with the medians and the 25th and 75th quantiles indicated, were computed and are shown in Fig. 13. Note that in each case, the interquartile ranges are non-overlapping, further reinforcing our belief that the performance gain of the DRF over the RF is statistically significant in the lower FAR region of interest.

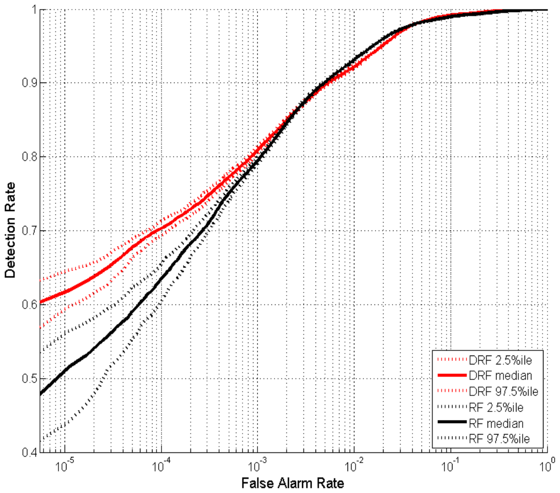

Fig. 12. The median ROC, 97.5 and 2.5 percentiles curves for the DRF (red, *m* = 1) and RF (black, *m* = 2) classifiers using the Hidden Signal data sets (*T1* and *J1*).

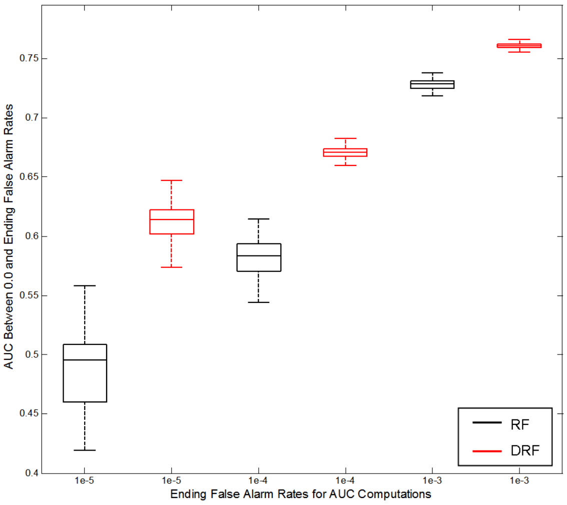

Fig. 13. Box plots of the AUC values for the DRF (red, *m* = 1) and RF (black, *m* = 2) classifiers using the Hidden Signal data sets (*T1* and *J1*).

4.4 Small Samples and Early Stopping

Both the RF and DRF are tree-based ensemble classification methodologies whose construction relies fundamentally upon the processes of bagging and random feature selection. In fact, the DRF shares many similarities with the RF and, consequently, shares many of its most noteworthy advantages. However, a critical exception is the parametric node-splitting process utilized by the DRF. The conventional RF approach to node splitting is accomplished via a univariate threshholding process that is optimized relative to some predetermined criterion (e.g., Gini impurity). Generally, parameters are not estimated during this process. In contrast, node splitting under the DRF regime is performed by building an LDA model at each node in an effort to determine an "optimal" linear decision boundary. For the two-class problem, this requires the estimation of the class specific mean vectors $\mu_k,\ k = 0, 1$ and common covariance matrix Σ at each node in the forest.

This is an important distinction between the two methods that manifests itself in several ways. In particular, the training of the lower portions of all DRF trees (near the leaf nodes) must contend with progressively sparser data sets. Specifically, the LDA models are based upon point estimates (e.g., maximum likelihood estimates) of the mean vectors and common covariance matrix. For a split dimension of $m \geq 1$, there are exactly $p = 3m + m(m-1)\big/2$ parameters to estimate at each node. Hence, as the value of m increases, the estimation problem becomes increasingly challenging at the more sparsely populated nodes in the forest. In severe cases, the common covariance matrix is not even estimable. This occurs exactly when a node is impure (i.e., both classes are represented) and the feature vectors within each class are identical. In these cases, the DRF splitting process defaults to a geometric method (i.e., the decision threshold is taken to be the perpendicular bisector of the chord connecting the sample means of the two classes). This tactic has proven relatively effective in practice, but the more subtle issue of small sample size parameter estimation remains.

Unfortunately, this is a common problem in statistics, and our study of the DRF methodology would be incomplete without investigating the role parameter estimation plays in its performance and reliability. A careful examination of the median ROC and the corresponding 97.5 and 2.5 percentile bands shown in Fig. 12 reveals behavior that may provide insight into this issue. Note that the distance between the percentile bands increases noticeably as the FAR drops from 10^{-4} to $5*10^{-6}$, suggesting a considerable increase in the variability of the experimental ROC curves over this extreme interval. It is our conjecture that a contributing factor to this phenomenon is the sparseness of the data in the lower nodes of the DRF trees, which leads to greater instability in the parameter estimates.

This behavior is even more pronounced for the RF, appearing at first glance to contradict the above conjecture. Though its underlying cause is not entirely clear, this contradictory behavior may be at least partially explained by the fact that each DRF node splitting decision utilizes two sources of information: the node data and the LDA model. The model may exert a dampening influence on the impact of the data, reducing variability at the leaf nodes and thus reducing variability in the ROC at low FAR values. In contrast, the RF is driven entirely by the data and hence may prove more vulnerable to data variation and sparseness.

In any case, enriching the data at the lower DRF nodes in an effort to improve parameter estimation and ultimately enhance performance (e.g., increase the median detection rate while reducing variability) is a challenge we would like to address. As a first attempt in this direction, we considered the optimal DRF ($m = 1$) for the Hidden Signal Detection application problem first described in Section 4.1. In this case, the LDA model building exercise reduces to estimating three univariate Gaussian parameters. We have observed in our studies that the number of samples used to estimate these parameters near the leaf nodes is frequently very small – less than 10 in many cases. To more fully explore this issue, suppose we require that at least $n = 30$ samples be used to estimate the LDA parameters at any node in the forest. *What is the impact of this constraint on the detection performance?*

To address this question, we incorporated an early stopping criterion into our methodology whereby tree nodes continue to successively split until either purity is achieved or the number of node samples drops below a prescribed value, *n*. This version of the DRF methodology is called “early stopping” DRF and is denoted by DRF-ES.

For the following experiment, we once again randomly generated a collection of 101 forests, trained on *T1* and tested on *J1*, for both the DRF and DRF-ES methodologies. Figure 14 presents the median ROC curves and the corresponding 97.5 and 2.5 percentile bands for the standard DRF and the DRF-ES with $n = 30$. We observed that for FAR values less than approximately $5*10^{-2}$, the DRF-ES classifier significantly outperforms the conventional DRF classifier. Figure 15 shows the corresponding box plots of the AUC values for FAR cutoff levels of 10^{-3}, 10^{-4}, and 10^{-5}. Over each of these intervals, the interquartile ranges are non-overlapping, reinforcing the statement that the performance gain of the DRF-ES over the DRF is significant in this case (m = 1).

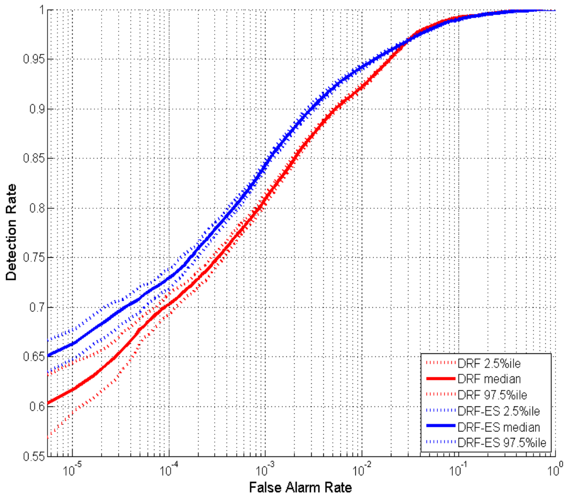

Fig. 14. The median ROC, 97.5 and 2.5 percentiles curves for the DRF (red) and DRF-ES (blue) classifiers for *m* = 1 using the Hidden Signal data sets (*T1* and *J1*).

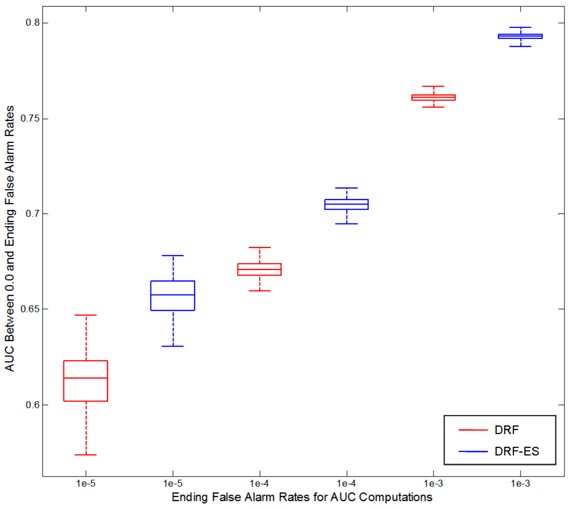

Fig. 15. Box plots of the AUC values for the DRF (red) and DRF-ES (blue) classifiers at m = 1 using the Hidden Signal data sets (*T1* and *J1*).

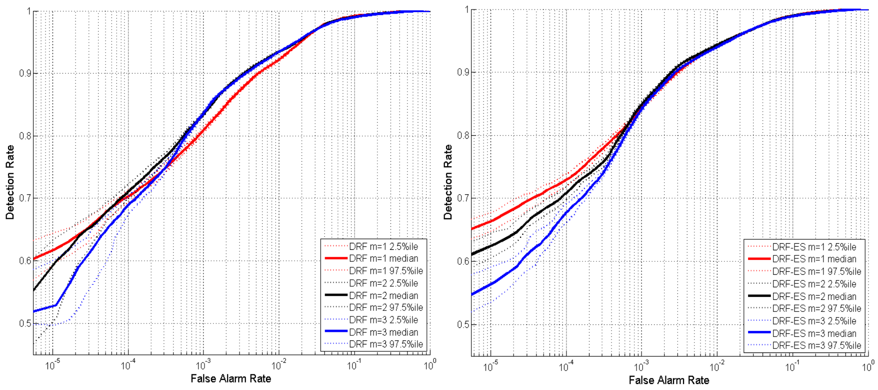

Fig. 16. Median ROC curves with 97.5 and 2.5 percentile bands for DRF (left) and DRF-ES (right) for the Hidden Signal Detection problem at m = 1, 2, 3 (*T1* and *J1*).

We then extended these studies and compared the performance of the DRF and DRF-ES classifiers for higher split dimensions of $m = 2$ and 3. Figure 16 shows the median ROC curves with percentile bands for the DRF and DRF-ES applied to the Hidden Signal Detection data sets. Note that in both cases the performance degrades as m increases, but the DRF-ES curves exhibit a “nesting” behavior (i.e.,

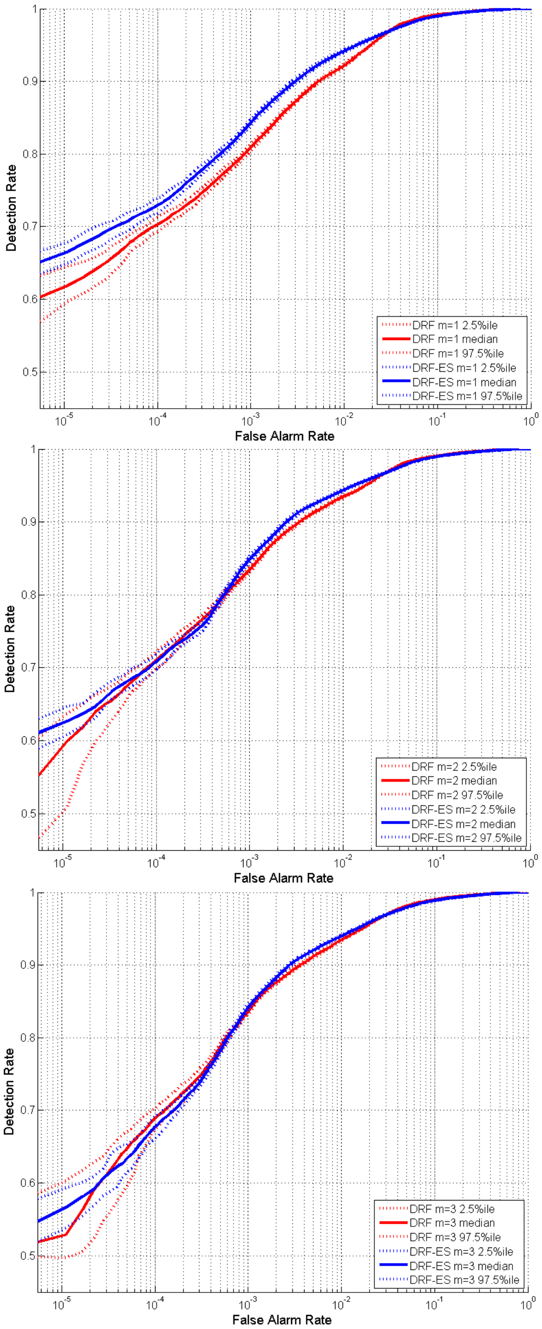

Fig. 17. Median ROC curves with 97.5 and 2.5 percentile bands for the DRF (red) and DRF-ES (blue) for *m* = 1, 2, 3 (numbered top to bottom).

they have similar shape) that suggests greater performance stability relative to the DRF. Specifically, we see that the DRF-ES produces narrower percentile bands that are non-overlapping, while those of the DRF are wider and repeatedly cross. In Fig. 17, the DRF and DRF-ES median ROC curves are plotted head-to-head, together with their 97.5 and 2.5 percentile bands, for each value of m. Note that the performance advantage enjoyed by the DRF-ES degrades as m increases from 1 to 2, evaporating entirely when m equals 3. In other words, as the number of parameters increases while holding the maximal sample size $n = 30$ constant, the DRF-ES performance gains are eroded. This behavior is consistent with our earlier conjecture that parameter estimation based upon sparse data, combined with increasing model dimensionality, adversely affects the performance and stability of the DRF methodology. However, it remains likely that the true mechanisms underlying these behaviors are quite complex and defy simple explanation. Issues such as model choice, misspecification, etc., may all be contributing factors.

4.5 Expected Cost

As we discussed in Section 1, information analysis systems are constructed for the purpose of compiling large stores of (potentially multi-source) data to support informed analysis and decision-making. Though many algorithms in the classification field are designed to minimize the expected overall error in class predictions, it is common for real-world detection problems to be inherently associated with unequal costs for false alarms and miss detections (e.g., the Hidden Signal Detection application). Thus, a natural performance metric that quantifies the expected cost of an incorrect decision in such cost-sensitive applications is given by:

$$EC = p(+) \cdot (1 - DR) \cdot c(miss) + p(\text{-}) \cdot FAR \cdot c(falsealarm) \tag{8}$$

where DR is the detection rate, $p(\cdot)$ is the prior probability for each class, and $c(\cdot)$ is the cost for each type of error. To enable visualization of the general behavior of this metric, Drummond and Holte developed “cost curves” that express expected cost as a function of the class priors and costs [7]. Specifically, cost curves plot the expected cost (normalized by its maximum value) versus the probability cost function (*PCF*), which is given by

$$PCF = \frac{p(+)\cdot c(miss)}{p(+)\cdot c(miss) + p(-)\cdot c(falsealarm)}. \tag{9}$$

Assuming equal priors, *PCF* is small when the cost of false alarms is large relative to that of missed detections. In the Hidden Signal Detection application, the cost of a false alarm is considered to be at least 100 times more costly than a missed detection, making classifiers whose cost curves are lower at small values of *PCF* (e.g., *PCF* < 0.01) more desirable. In Fig. 18, we have plotted the median, 2.5 percentile, and 97.5 percentile cost curves for the RF, DRF, and DRF-ES. We immediately observe that the DRF and DRF-ES appear to be significantly more effective than the RF, with DRF-ES achieving the smallest expected cost across the low PCF range of interest.

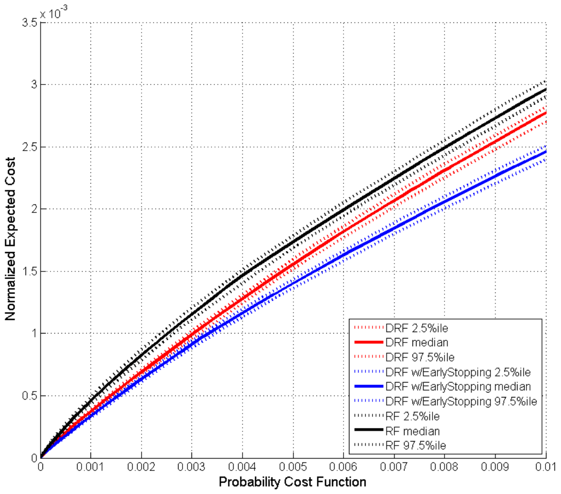

Fig. 19. Cost curves plotted for RF (black), DRF (red) and DRF-ES (blue) over the PCF range of interest. The discriminant-based classifiers outperform the RF over this range, with the DRF-ES achieving the lowest cost over all.

5 Conclusions

The empirical results presented in Section 4 provide strong evidence that the Discriminant Random Forest and its ES variant produce significantly higher detection rates than the Random Forest over the low FAR regions of interest. The superior strength and diversity of the trees produced by the DRF further support this observation. In addition, this methodology appears to be more successful in the detection of weak signals and may be especially useful for applications in which low signal-to-noise ratios are typically encountered.

We have also found that our methodology achieves far lower prediction errors than the RF when high-dimensional feature vectors are used at each tree node. In general, we expect the performance of any forest to decline as the number of features selected at each node approaches the cardinality of the entire feature set. Under these conditions, the individual base classifiers that compose the forest are nearly identical, negating many of the benefits that arise from an ensemble-based approach. In such cases, the only variation remaining in the forest is due to the bagging of the input data. Though we observed the expected performance degradation for both forest methodologies at extremely high split dimensions, the effect was far less severe for the discriminant-based approach. This result suggests a versatility and robustness in the DRF methodology that may prove valuable for some application domains.

Our investigation into the effect of sparse data revealed that an early stopping strategy might help mitigate its impact on classification performance, ultimately increasing the detection rate over the lower FAR regions. This advantage, however, is diminished as split dimensionality increases. Though we did not thoroughly explore the sensitivity of the DRF performance to the early stopping parameter, n, we conjecture that increasing this parameter in response to increases in split dimensionality would be counterproductive. Such a strategy would eventually eliminate the fine-grained (i.e., small scale) class distinctions that are critical for effective classification. However, at split dimension $m = 1$, the DRF enjoys a significant performance improvement in low FAR regions via early stopping. In fact, for our applications, the DRF-ES at $m = 1$ outperformed the DRF over all dimensions.

Overall, our empirical studies provided considerable evidence that the DRF is significantly more effective than the RF over low FAR regions. Although computational efficiency may be adversely impacted by the more complex node-splitting of the DRF at extremely high dimensions, its peak performance is typically achieved at much lower dimensions where it is more efficient than the RF with respect to memory and runtime.

The behavior of the Discriminant Random Forest methodology is compelling and hints at complex internal mechanisms that invite further investigation. However, we have found statistically significant evidence supporting this technique as a highly robust and successful classification approach across diverse application domains.

Acknowledgments This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

1. L. Breiman, “Bagging Predictors”, Machine Learning, vol. 26, no. 2, pp. 123-140, 1996.
2. L. Breiman, “Using Adaptive Bagging to Debias Regressions”, Technical Report 547, Statistics Dept. UC Berkeley, 1999.
3. L. Breiman, “Random Forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
4. L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Chapman and Hall, 1984.
5. M. R. Chernick, Bootstrap Methods, A Practitioner's Guide. New York: John Wiley and Sons, Inc., 1999.
6. T. Dietterich, “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization,” Machine Learning, pp. 1-22, 1998.
7. C. Drummond and R. Holte, “Explicitly Representing Expected Cost: An Alternative to ROC Representation”, in Proc. of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000.
8. R. O. Duda, P. E. Hart, and D. H. Stork, Pattern Classification, 2nd edition, New York: Wiley Interscience, 2000.

9. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York: Chapman and Hall/CRC, 1993.

10. T. Fawcett, “ROC Graphs: Notes and Practical Considerations for Researchers”. Technical Report, Palo Alto, USA: HP Laboratories, 2004.

11. R. A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” Annals of Eugenics, vol. 7, pp. 179-188, 1936.

12. T. K. Ho, “Random Decision Forest”, in Proc. of the 3rd International Conference on Document Analysis and Recognition, pp. 278-282, 1995.

13. T. K. Ho, “The Random Subspace Method for Constructing Decision Forests”, IEEE Trans. On Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.

14. M. Loeve, Probability Theory II (Graduate Texts in Mathematics), 4th edition. New York: Springer-Verlag, 1994.

15. K. Mardia, J. Kent, J. Bibby, Multivariate Analysis. Academic Press, 1992.

16. G. J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition. New York, Wiley-Interscience, 2004.

17. J. J. Rodriguez, L. I. Kuncheva, et al., “Rotation forest: A New Classifier Ensemble Method,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 10, pp. 1619-1630, 2006.

18. *The Physical Protection of Nuclear Material and Nuclear Facilities*, IAEA INFCIRC/225/Rev.4 (Corrected).